MatFusion: A Generative Diffusion Model for SVBRDF Capture

Sam Sartor
College of William & Mary
Williamsburg, USA
slsartor@wm.edu

Pieter Peers
College of William & Mary
Williamsburg, USA
ppeers@siggraph.org

Figure 1: SVBRDF diffusion estimates visualized with integrated normal maps and global illumination for four different spatially varying materials captured by: a colocated flash photograph (1st and 2nd), a photograph captured under uncontrolled natural lighting (3rd), and a flash/no-flash image pair (4th).

ABSTRACT

We formulate SVBRDF estimation from photographs as a diffusion task. To model the distribution of spatially varying materials, we first train a novel unconditional SVBRDF diffusion backbone model on a large set of 312,165 synthetic spatially varying material exemplars. This SVBRDF diffusion backbone model, named MatFusion, can then serve as a basis for refining a conditional diffusion model to estimate the material properties from a photograph under controlled or uncontrolled lighting. Our backbone MatFusion model is trained using only a loss on the reflectance properties, and therefore refinement can be paired with more expensive rendering methods without the need for backpropagation during training. Because the conditional SVBRDF diffusion models are generative, we can synthesize multiple SVBRDF estimates from the same input photograph from which the user can select the one that best matches the users' expectation. We demonstrate the flexibility of our method by refining different SVBRDF diffusion models conditioned on different types of incident lighting, and show that for a single photograph under colocated flash lighting our method achieves equal or better accuracy than existing SVBRDF estimation methods.

CCS CONCEPTS

• **Computing methodologies** → **Reflectance modeling**.

KEYWORDS

SVBRDF, Diffusion, Appearance Modeling

ACM Reference Format:
Sam Sartor and Pieter Peers. 2023. MatFusion: A Generative Diffusion Model for SVBRDF Capture. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3610548.3618194

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0315-7/23/12…$15.00
https://doi.org/10.1145/3610548.3618194

1 INTRODUCTION

Reproducing the visual appearance of real-world spatially varying materials is a challenging research problem that requires balancing multiple competing goals such as ease of capture, robustness, accuracy of the reproduction, and suitability for post-production editing. The most promising recent solutions leverage machine learning to produce Spatially Varying Bidirectional Reflectance Distribution Function (SVBRDF) parameter maps that correspond to one or more photographs of the target material. These methods are convenient and can produce plausible SVBRDFs. However, SVBRDF modeling is inherently ambiguous as multiple parameter combinations can explain the (underconstrained) appearance observations of the material, and there is no recourse when the inferred property maps fail to reproduce plausible material properties; there are typically no additional hyper-parameters that can be tuned to produce alternative solutions. Furthermore, these machine learning based methods are trained for a specific type of incident lighting, and modifying the input lighting often requires a lengthy retraining step and an appropriate corresponding loss.

Inspired by recent successes in using diffusion models [Karras et al. 2022; Rombach et al. 2022; Song et al. 2021b] for image synthesis tasks such as image restoration [Dhariwal and Nichol 2021; Ho et al. 2020, 2022], super-resolution [Kadkhodaie and Simoncelli 2021; Saharia et al. 2023], and image-to-image translation [Saharia

et al. 2022; Sasaki et al. 2021] we formulate SVBRDF estimation as a diffusion task. Existing diffusion based image processing methods rely on pre-trained large scale image diffusion models to sample the distribution of natural images. However, the distribution of SVBRDFs differs significantly from natural images. We therefore introduce a novel generative diffusion model geared towards spatially varying materials. We introduce an unconditional backbone diffusion model, named MatFusion, that synthesizes SVBRDF parameter maps (i.e., diffuse and specular albedo, specular roughness, and normals). We leverage ConvNeXt blocks [Liu et al. 2022] instead of the typical Residual blocks [He et al. 2016] commonly used in diffusion models to increase the number of activations without increasing the parameter count to better model the 10 SVBRDF channels (versus 3 for images). Furthermore, training diffusion models typically requires a significantly larger training set than conventional convolutional neural networks. To support training an SVBRDF diffusion model, we supplement the INRIA synthetic SVBRDF dataset [Deschaintre et al. 2018] with a new training set constructed from 1,877 synthetic SVBRDFs, that after augmentation with a novel mixing strategy, together with the INRIA dataset, grows to 312,165 unique training exemplars. Building on the MatFusion backbone, we also introduce *three* conditional refinements that differ in their input: the classic colocated camera-flash image, a photograph under uncontrolled natural lighting, and a flash/no-flash image pair (Figure 1). By changing the seed, all three models can produce a variety of candidate SVBRDF replicates, from which the SVBRDF that best matches the user's expectation can be selected. Our backbone diffusion network is trained using only SVBRDF parameter losses (i.e., without a rendering loss), and thus no backpropagation through a differentiable renderer is needed. This allows us to train the conditional diffusion network on input images that contain a more complete characterization of the surface reflectance by integrating the normal maps and accounting for indirect lighting within the material. While such indirect lighting does not contribute significantly for backscatter surface reflectance, it does impact the visual appearance significantly for more complex lighting conditions (such as natural lighting).

We demonstrate the efficacy of finetuning the MatFusion backbone and show that the conditional diffusion networks produce plausible SVBRDFs, and in case of colocated flash lighting, with equal or better quality than existing methods.

In summary, our contributions are:

(1) MatFusion: a backbone k-diffusion model that generates 10 channels of reflectance properties;
(2) three conditional SVBRDF diffusion models refined from the MatFusion backbone using a novel direct conditioning strategy; and
(3) a training set of 312,165 unique synthetic SVBRDFs.

2 RELATED WORK

We focus the discussion of related work on learning-based generative and inference methods for modeling SVBRDFs.

Direct Inference Methods. Estimating spatially varying material parameters from a single photograph is a difficult problem. Leveraging advances in neural networks, Li et al. [2017] and Ye et al. [2018] demonstrate plausible SVBRDF capture from a single photograph under unknown natural lighting, albeit restricted to a predetermined class of materials (e.g., metals, plastics, etc.) Deschainte et al. [2018] introduced the de-facto standard training set of approximately 200,000 synthesized SVBRDFs to train an inference network, using a novel render loss, that estimates the SVBRDF property maps from a single photograph lit by a colocated flash light. Subsequent work further improved the inference accuracy by exploring novel architectures and loss functions [Guo et al. 2021; Li et al. 2018; Sang and Chandraker 2020; Vecchio et al. 2021; Zhou and Kalantari 2021] or supporting multiple input photographs [Deschaintre et al. 2019; Ye et al. 2021]. Martin et al. [2022] capture SVBRDFs, albeit without specular albedo, from outdoor photographs that include ambient occlusion effects. All of the above methods are trained for a specific input lighting condition; it is unclear to what degree the architecture and loss are tuned to the expected lighting, and significantly changing the lighting condition during capture would require retraining the network from scratch. In contrast, our method builds on an unconditional SVBRDF diffusion backbone, trained independently from the incident lighting, which can serve as a basis for conditional finetuning. Furthermore, all the above methods produce a single result per photograph, and offer no strategies for producing alternative estimates that can better explain the appearance.

Iterative Inference Methods. In contrast to direct inference methods that directly produce the target material property maps, iterative inference methods perform an online optimization to minimize a rendering loss with respect to the captured photograph. Gao et al. [2019] and Guo et al. [2020b] perform the optimization in a learned space modeled by an auto-encoder and a GAN respectively. In both cases, the lighting condition is only considered during the online optimization process, and the space of SVBRDFs is lighting agnostic. Hence, these methods could in theory be applied to different lighting conditions. However, neither method provides an interface for directing the optimization process to different plausible SVBRDFs. Furthermore, both methods tend to suffer from over-fitting, resulting in burned-in highlights in the diffuse albedo maps. Zhou and Kalantari [2022] and Fischer and Ritschel [2022] combat overfitting by combing direct inference and optimization-based methods using meta-learning. While this greatly improves the quality, the resulting trained networks are lighting specific. Our method is also iterative, but unlike the above methods, we do not minimize a render loss function, but instead solve a denoising differential equation. Unlike prior iterative methods, our method can produce different replicate SVBRDFs by changing the input seed.

Generative Methods. Aittala et al. [2016] extend parametric texture synthesis to replicate the spatially varying appearance of a mostly stationary material from a single flash lit photograph of an exemplar material. Similarly, Wen et al. [2022] train a GAN to model the appearance from a photograph of a stationary material. Henzler et al. [2021] employ a convolutional neural network, conditioned on a latent code from a learned space, to convert a random noise field into a random non-repeating field of BRDFs that match the appearance of a flash-lit photograph of a stationary material. Inspired by MaterialGAN [Guo et al. 2020b], Zhou et al. [2022] and Hu et al. [2022a] introduce tileable material GANs that allow for spatial control through an additional guidance image. While these networks can produce some stochastic variations

around the expected value, they do not effectively sample the distribution conditioned on the input image. In contrast our method samples the conditional SVBRDF distribution that better adheres to the input material's appearance. An alternative strategy to directly synthesizing the SVBRDF property maps, is to generate a procedural model [Guerrero et al. 2022; Hu et al. 2022b; Shi et al. 2020]. The parameters of such procedural models can be matched to the appearance of an exemplar in a photograph [Guo et al. 2020a]. However, current procedural methods are limited to specific material classes.

3 SVBRDF DIFFUSION MODEL

Preliminaries. We model the appearance of a planar spatially varying material by an SVBRDF, where each surface point's reflectance is modeled by a microfacet BRDF with a GGX distribution [Walter et al. 2007] parameterized by its diffuse albedo, specular albedo, and monochrome specular roughness. In addition, we model the local surface variations by a normal map.

MatFusion. We first model the distribution of SVBRDFs using an unconditional diffusion model, named MatFusion, that we will subsequently refine based on the capture conditions. The basic observation of diffusion modeling is that adding noise to a signal (e.g., image) is a destructive process, and hence the process of removing noise must therefore be generative. In the limit, an entirely synthetic signal can be generated by starting from pure random Gaussian noise, and iteratively denoising the signal [Ho et al. 2020]. Formally, the goal of a generative model is to sample a random variable according to a target data distribution $x_0 \sim p_{\text{data}}$. In a diffusion model, we consider a sequence of related random variables $x_{1,2,\ldots,T}$ where each subsequent variable is increasingly more noisy until x_T is indistinguishable from pure Gaussian noise:

$$p(x_t|x_0) = \mathcal{N}\left(x_0, \sigma_t^2\right), \tag{1}$$

with $\sigma_t > \sigma_{t-1}$. The diffusion process itself repeatedly samples $p(x_{t-1}|x_t)$ starting with $t = T$ and ending when $t = 0$ [Ho et al. 2020; Song et al. 2021a]. This differs from a traditional generator (e.g., GAN) that samples x_0 directly. Song et al. [2021b] formulate diffusion as a differential equation that maintains the distribution p as x evolves over time. The change in x with time t is then[1]:

$$\mathbf{d}x = -\dot{\sigma}(t)\sigma(t)\nabla_x \log p(x;\sigma(t))\mathbf{d}t, \tag{2}$$

where $\dot{\sigma}(t)$ denotes the time-derivative of $\sigma(t)$. $\nabla_x \log p(x;\sigma(t))$ is also called the score function: a vector that points towards the highest density of probable signals. The differential denoising equation can then be solved by taking discrete time-steps to evolve the solution (e.g., using an Euler method) using Equation (2). To compute the score function, we define a neural denoising network $D_\theta(x_t;t)$ that minimizes the expected error on samples drawn from p_{data} for every σ_t. To avoid that the inputs of D_θ grow with increasing σ_t, it is standard practice to normalize the estimate x_t by $\sqrt{1+\sigma_t^2}$. Denoting the normalization factor of x_t as a, abstracts the network input y as $ax + bn$ s.t. $a^2 + b^2 = 1$, where n is Gaussian distributed noise [2]. Karras et al. [2022] introduced a robust diffusion variant,

[1]We assume no time-dependent signal scaling, i.e., s(t) = 1.

[2]a and b in this case correspond to $\sqrt{\bar{\alpha}}$ and $\sqrt{1-\bar{\alpha}}$ in [Ho et al. 2020].

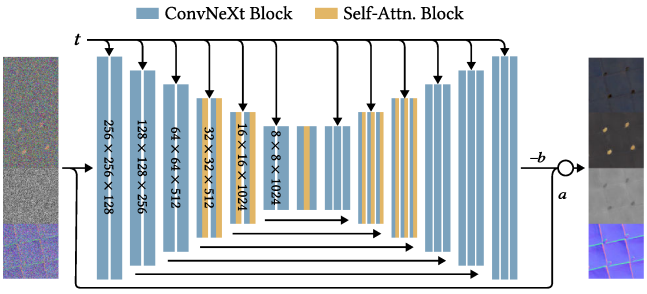

Figure 2: Summary of the MatFusion architecture.

named k-diffusion, that instead of estimating the noise as in prior diffusion models, estimates the "velocity" $an - bx$ (note the swapped position of n and x and change of sign for the second term) such that the denoising network D_θ minimizes the loss function:

$$\mathbb{E}_{x\sim p_{\text{data}}}\,\mathbb{E}_{n\sim\mathcal{N}(0,1)}\,\|D_\theta(y;t) - (an - bx)\|_2^2, \tag{3}$$

This allows us to estimate both the expectation of noise and signal with equal ease by leveraging that $a^2 + b^2 = 1$:

$$\mathbb{E}_n \approx by + aD_\theta(y;t), \tag{4}$$

$$\mathbb{E}_x \approx ay - bD_\theta(y;t). \tag{5}$$

Note that depending on a (which depends on $\sigma(t)$), the output of the neural network D_θ varies from an estimate of the signal x to and estimate of the noise n when $t \to 0$.

Architecture. In this paper we follow the normalization and sampling schedule (i.e., $\sigma(t)$) from [Ho et al. 2020], but use the k-diffusion loss function for D_θ. Our architecture for D_θ is inspired by Dhariwal et al. [2021]'s ImageNet-256 U-net architecture with 6 resolutions for the encoder and decoder (Figure 2). To accommodate for the larger number of channels (10 for SVBRDFs vs. 3 for images), we employ a $3 \times 3 \times 10 \times 128$ convolution kernel to transform the 10 input channels into 128 features. We replace the Residual convolution blocks with ConvNeXt blocks [Liu et al. 2022] to increase the number of activations for the same number of parameters; we argue that the higher channel count benefits from more activations. We follow DDIM [Song et al. 2021a] and encode t as a 512-length feature (using Fourier embedding and a 2-layer MLP) and pass it to each ConvNeXt block as a dense residual layer between the 7×7 convolution and the first depth-wise convolution. Similar to DDIM, all layers use a group norm with 32 groups, and the 32 and 16 resolution layers include self-attention blocks (with 8 heads) after each ConvNeXt block, as well as an additional attention-layer at the bottleneck. We follow the method of Rabe et al. [2021] to reduce the memory overhead of the attention layers during training.

Conditional SVBRDF Diffusion Model. In order to recover a plausible SVBRDF from a photograph, we need to make the SVBRDF diffusion backbone network conditional on the photograph. One possible strategy to condition the neural network D_θ on additional input images is by concatenating them to the input noise [Saharia et al. 2022; von Platen et al. 2022]. However, this would require retraining the diffusion network from scratch which is very costly. Vonyov et al. [2022] perform sketch-guided text-to-image diffusion by backpropagating the loss over the condition and an inverse

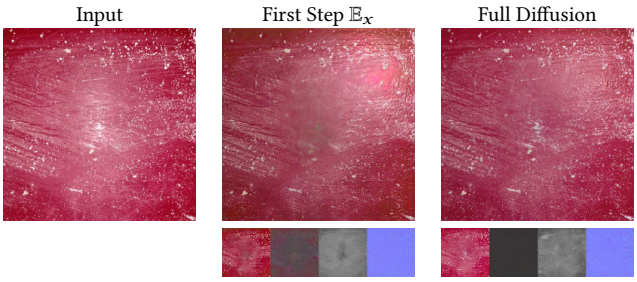

Figure 3: For the first diffusion step, the denoising neural network D_θ fully relies on the input photograph (left) and acts as a direct inference network (middle). However, in contrast to direct inference, a diffusion model iteratively improves the estimate (right) by reducing burn-in, adding detail in the normal map, and improving diffuse-specular separation.

mapping from the diffusion output to the condition. In the context of SVBRDFs, this would be akin to driving the diffusion process by the render error, risking burn-in artifacts. Recently, Zhang and Agrawala [2023] showed that an existing unconditional diffusion model can be conditioned by adding zero-initialized dense layers to each skip connection, and providing them the outputs of a parallel control network trained on the conditional task.

Inspired by Zhang and Agrawala [2023], we expand the *input head* with k additional features with both weights and bias initialized with zeros (i.e., yielding an initial convolution kernel of $3 \times 3 \times (10 + k) \times 128$, and where $k = 3N$, and N is the number of condition input photographs). Next, we *finetune* the backbone model for the target type of input photographs (unlike direct concatenation which requires retraining from scratch). Compared to ControlNet, our approach is easier to implement and incurs less overhead as we do not need an additional control network (we only expand the input head) at the cost of "polluting" the original diffusion network.

Relation to Direct Inference. When the k-diffusion model is conditioned on a photograph c of the target material, the model subsumes direct inference methods. At $t = T$, the signal $y = ax + bn$ is purely Gaussian noise (i.e., $a \sim 0$), and hence $D_\theta(y|c; t)$ mostly relies on the condition c to estimate the velocity (i.e., $an - bx \sim x$). For all practical purposes, we can ignore the noisy input at $t = T$, and thus the *expectation* $\mathbb{E}_x$ computed from the estimate of D_θ (Equation (5)) closely mimics the behavior of a direct inference method. However, unlike direct inference methods, diffusion only takes a small step towards the estimate and continues to improve the result in subsequent steps. Figure 3 demonstrates that the expectation from the first diffusion step is similar to the result of a direct inference method; note all SVBRDF property maps shown in this paper are ordered as: diffuse albedo, specular albedo, roughness, normal map. This initial estimate often exhibits burn-in, bended normals and missing details, and imprecise diffuse-specular separation, which are reduced in subsequent diffusion steps.

4 TRAINING DATA

The MatFusion backbone model has 256M parameters, hence, training such a model requires a large and diverse training set. Deschaintre et al. [2018] augment 150 synthetic SVBRDFs to 199,068 training exemplars by randomly perturbing parameters, scaling/rotating the exemplars, and taking convex combinations. However, since the dataset is augmented from only 150 SVBRDFs, the texture diversity is limited and insufficient to train our MatFusion backbone model. To mitigate this issue, we collected and augment 307 additional synthetic SVBRDFs from https://polyhaven.com and 1,570 additional synthetic SVBRDFs from https://ambientcg.com.

The 307 SVBRDFs from Polyhaven are CC0 licensed and each contains a unique diffuse albedo map, normal map and roughness map at $2k$ resolution. Polyhaven's SVBRDFs do not come with a specular albedo. We therefore assign a homogeneous specular albedo uniformly sampled in $[0.04, 0.08]$. The 1,570 SVBRDFs from AmbientCG are also CC0 licensed, and all contain unique albedo, specular roughness, and normal maps at $2k$ resolution. 274 SVBRDFs also contain a metalness map. A homogeneous specular albedo is assigned (uniform random in $[0.04, 0.08]$) plus albedo times metalness (if available). The diffuse albedo is set to the albedo (scaled by one minus metalness if available).

For each of the 1,877 SVBRDF maps we randomly crop 16 square areas, each from from a random position, rotation, and size (between 512 and 1,400 pixels fully contained within the original maps). Each cropped map was bilinearly resized to 512×512 resolution, yielding a total of 30,032 basis SVBRDFs. To further diversify the roughness maps, we randomly select 6,000 basis SVBRDFs, and blend their roughness maps with procedurally generated maps. We employ a randomly initialized dense neural network that transforms each pixels' (diffuse + specular) albedo and height (obtained by integrating the normal map [Quéau et al. 2018]) to a procedural roughness value; see the supplemental material for more details. Note, the randomly initialized network is not optimized and it serves as a random non-linear transformation of albedo and height to roughness.

To better mimic that real-world materials are often formed by piece-wise constant combinations of different basis materials (e.g., metal and rust), we create 83,065 additional piece-wise constant mixtures from both the 199,068 INRIA SVBRDFs and the 30,032 basis SVBRDFs. For 66% we mix two randomly selected SVBRDFs without replacement (i.e., each SVBRDF is only used in one mixture material), and three SVBRDFs for the remaining 34%. We use a randomly initialized dense neural network (detailed in the supplemental material) that transforms each pixels' (diffuse + specular) albedo and height into a one-hot selection weight (for each of the two/three source SVBRDFs). Similar as for the roughness generator, the randomly initialized network is not optimized and it serves as a random non-linear transformation and thresholding step. To avoid unnatural hard edges, we perform the mixing on 2× bilinearly upsampled randomly selected 288×288 crops from the INRIA or basis SVBRDFs, and after mixing, (average) downsample again to 288×288 resolution.

Combining the INRIA training set (199,068 at 288×288 resolution), our basis SVBRDF set (30,032 at 512×512 resolution), and the mixture set (83,065 at 288×288 resolution) yields our final training

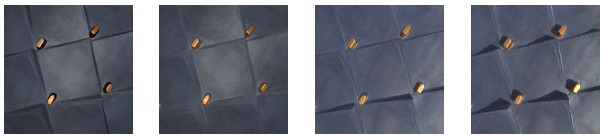

Figure 4: Global illumination transport within the spatially varying material is negligible for a colocated camera-light setup. However, under natural lighting, the effects are significant (i.e., self-shadowing and ambient occlusion).

set with 312,165 training exemplars. In addition, we created a test set of 50 materials that consists of a selection of 31 diverse materials from the Deep Inverse Rendering [Gao et al. 2019] test set, 6 materials from the look-ahead meta-learning [Zhou and Kalantari 2022] test set, 11 from Polyhaven, and 2 from AmbientCG. None of the test materials are included in the training set.

5 RESULTS

Implementation. We implemented MatFusion in FLAX [Heek et al. 2023] and train it for 50 epochs using the full 312,165 SVBRDF training set (cropped to 256×256 resolution) using the AdamW optimizer [Loshchilov and Hutter 2019] with a batch size of 32, a learning rate of 2×10^{-5} (with a 100,000 iteration warmup), and EMA weights [Song and Ermon 2020] on 4 Nvidia A40 GPUs with 48GB of memory. Training took approximately 255 hours.

We train three conditional variants of MatFusion. All three are finetuned for 19 epochs on MatFusion using the full SVBRDF training set using the same optimizer and hyperparameters. Training took approximately 102 hours on 4 Nvidia A40 GPUs, or 2.5× faster than training MatFusion from scratch. The three variants differ in the expected lighting in the input condition photograph: COLOCATED flash lighting, FLASH/NO-FLASH, and NATURAL lighting. The COLOCATED variant is trained on synthetic photographs rendered with direct illumination only, as indirect lighting is negligible for backscatter reflectance. However, indirect lighting significantly affects the appearance of spatially varying materials (Figure 4). Therefore, the NATURAL and FLASH/NO-FLASH variants are trained on images rendered with Blender's Cycles path-tracer with 32 samples per pixel with OpenImageDenoise using the height map as the material's geometry obtained by integrating the surface normals [Quéau et al. 2018]; we use the original normal maps to determine the shading normals. Natural illumination is modeled by randomly selecting and rotating an HDR environment map from 560 CC0 licensed HDR environment maps retrieved from https://polyhaven.com/hdris. For the FLASH/NO-FLASH variant, the *log* relative brightness ratio between the flash lighting and the environment lighting is randomly sampled between $\log(1/50)$ and $\log(3/2)$. Both the NATURAL and FLASH/NO-FLASH variants are trained on images rendered with a virtual camera with a focal length of 35mm (i.e., camera distance = exemplar size). The COLOCATED variant is trained for a variable camera distance (with matching FOV) sampled according to a $\frac{1}{2}\Gamma(2, 2)$ distribution (relative to the exemplar size), and we concatenate the per-pixel view vector as an additional input condition.

During inference, the differential equation is iteratively solved using the EulerA solver [Song et al. 2021b] in just 20 steps and with the guidance scale set to 1.

Selection. The conditional SVBRDF diffusion models take, besides the input photograph, also a normal distributed random field determined by a seed. By changing the seed, different replicates of the SVBRDF can be generated (Figure 5). The choice of the seed can impact the quality of the result. Therefore, we show results selected with one of the following three selection strategies:

(1) *Fixed seed*: the seed is fixed for all results.
(2) *Render error selection*: we render the generated SVBRDFs from 10 random seeds and select the one that minimizes the LPIPS error [Zhang et al. 2018] when rendered under the capture lighting conditions.
(3) *Manual selection*: a set of 10 SVBRDFs generated with different random seeds are presented and the user manually selects the SVBRDF that appears (subjectively) the most plausible.

We also experimented with optimizing the input random field on the render error, but found that this tends to produce burn-in of the specular highlight. While the majority of seeds do not produce burn-in, those that do are scattered through the whole space. Thus no matter the starting point, there is always a nearby point that produces burn-in which the optimization will inevitably drive the solution towards.

Synthetic Results. Figure 9 compares the estimated SVBRDFs, manually selected from 10 random seeds, for 6 selected synthetic materials for each of the three conditional diffusion models. For each material, we show two renderings under different point lights for each of the models and the reference. In general, the COLOCATED model produces the most consistent results due to the known lighting, although it sometimes fails to recover the specular reflectance on small features (e.g., the nob in the 2nd material) or produces unexpected texture variations (e.g., the center of the 6th material). The results from the NATURAL model exhibit a greater variability in accuracy, such as incomplete diffuse-specular separation (4th example), or underestimation of specular roughness (6th example). Nevertheless, the resulting SVBRDFs are still plausible, demonstrating the ability of MatFusion to recover the SVBRDFs of general spatially varying materials under unknown lighting. The FLASH/NO-FLASH model benefits from having an input without strong specular highlights (i.e., no-flash) to better recover the diffuse texture. On the other hand, due to the unknown relative brightness of the natural lighting versus the flash lighting, it sometimes underestimates either the diffuse albedo (e.g., 4th material) or the specular roughness (e.g., 3rd material). The FLASH/NO-FLASH model shows that MatFusion can be conditioned on more than one input.

Comparison to Prior Work. Figure 10 compares the COLOCATED variant for each of the three selection methods (*fixed seed*, *render error*, and *manual* selection) against the adversarial direct inference method of Zhou and Kalantari [2021] and the meta-learning look-ahead method of Zhou and Kalantari [2022] on synthetic SVBRDFs. Qualitatively, the COLOCATED model produces a more plausible appearance and the corresponding property maps appear "cleaner". These qualitative conclusions are supported by the average LPIPS

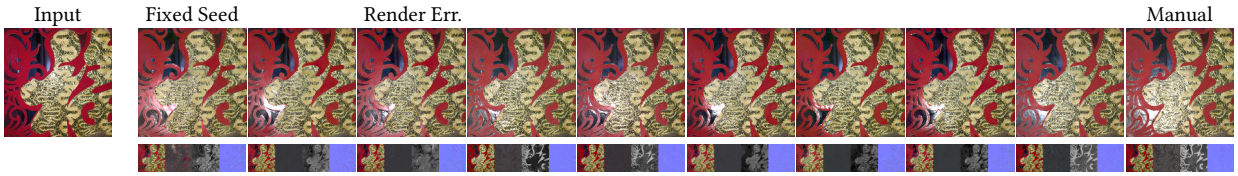

Figure 5: Changing the seed results in different SVBRDF replicates conditioned on the input photograph. For each replicate we show a rendering under a different lighting than the input photograph as well as the generated SVBRDF property maps. Also marked are the SVBRDF selection based on the render error with respect to the input lighting, as well as the manual selection of the (subjectively) most plausible SVBRDF.

Table 1: Quantitative comparison of average RMSE on the property maps and average LPIPS errors on 128 renders lit by a uniformly sampled point light on the hemisphere for the COLOCATED conditioned MatFusion model versus Zhou and Kalantari's [2021] adversarial inference method and Zhou and Kalantari's [2022] meta-learning look-ahead method.

	LPIPS	RMSE			
	Render	Diff.	Spec.	Rough.	Normal
Adversarial	0.2304	0.0439	0.0859	0.1358	0.0577
Adversarial (retrained)	0.2292	**0.0405**	0.0795	0.1276	0.0545
Look-ahead	0.2647	0.0591	0.0727	0.1424	0.0572
MatFusion (fixed seed)	0.2282	0.0427	0.0691	**0.1252**	0.0561
MatFusion (render err.)	0.2138	0.0440	**0.0657**	0.1282	0.0543
MatFusion (manual)	**0.2056**	0.0412	0.0666	0.1265	**0.0524**

Table 2: Achitecture ablation study of average RMSE on the property maps and average LPIPS render errors on 128 visualizations lit by a uniformly sampled point light, comparing the impact of using Residual convolution blocks versus ConvNeXt convolution blocks, and comparing the difference between using ControlNet and our direct conditioning.

	LPIPS	RMSE			
	Render	Diff.	Spec.	Rough.	Normal
ResNet+Control	0.2655	0.0525	0.0813	0.1536	0.0545
ConvNeXt+Control	0.2731	0.0517	0.0764	0.1428	0.0604
ResNet+Direct	0.2093	0.0432	0.0682	**0.1055**	0.0528
ConvNeXt+Direct	**0.2056**	**0.0412**	**0.0666**	0.1265	**0.0524**

[Zhang et al. 2018] render error listed below. We render each exemplar over a set of 128 randomly selected point lights on the hemisphere (with a radius of 2.41 units to match the training (and thus offer a best case evaluation) of Zhou and Kalantari [2021; 2022]), as well as in Table 1 for manual selection on the whole test set of 50 materials. We argue that a perceptual render error is the best metric for comparing the different methods as different maps can produce similar material appearances. For completeness, Table 1 also lists the RMSE errors over the SVBRDF property maps. We also include a comparison to Zhou and Kalantari's adversarial direct inference method retrained using our training set. MatFusion is a generative model which does not guarantee pixel-perfect alignment, which can result in sometimes a larger error on texture-rich property maps (e.g., 6th row) or unobserved properties (e.g., 2nd row). However, qualitatively, these property maps include fine details, albeit not perfectly aligned with the reference. In contrast, the look ahead-method of Zhou and Kalantari [2022] produces normal maps with little detail, resulting in a low error, but distributed over the whole map. Figure 10 also demonstrates that the render error selection can provide a good match (e.g., 1st and 5th row), but it can also overfit (e.g., 3rd row).

Real-world Validation. Figure 6 and Figure 7 demonstrate that MatFusion generalizes well to real-world captures. The results in Figure 6 are manually selected from 10 random seeds and validated on the materials captured by Guo et al. [2020b] which also contain reference photographs captured under different lighting conditions. Our results are visually closer to the reference than the adversarial direct inference method of Zhou and Kalantari [2021], and the look-ahead method of Zhou and Kalantari [2022]. Our method suffers less from specular burn-in (1st example) and overfitting normal detail to specular highlights in the input (2nd and 3rd example).

The materials in Figure 7 are captured in-the-wild by us using a *Pixel 5a* cell phone, and we manually select the most plausible SVBRDFs. Note that these images are captured under unknown natural lighting, and due to the uncontrolled nature of the capture conditions, no reference photographs under different lighting conditions are available. Nevertheless, the SVBRDF property maps nicely separate diffuse and specular, and the renderings plausibly capture the appearance from the input photographs.

Ablation Study. We perform an ablation study to justify the design decisions with respect to the architecture of MatFusion (Table 2). We validate both the impact of using Residual versus ConvNeXt convolutional blocks and using ControlNet versus direct conditioning. For all models we compute the average RMSE on the property maps and average LPIPS error on renders under the same set of random point lights for each of the 50 test materials. From Table 2, we observe that ConvNeXt layers slightly outperform Residual convolutional blocks on LPIPS error and $\sim 5\%$ better on RMSE on the albedos; the lower roughness error for ResNet is due to a few outlier materials. Furthermore, direct conditioning outperforms ControlNet on all metrics, while training time is similar for both, except that ControlNet requires significantly more memory resources. We posit that the difference in performance is due to ControlNet only receiving indirect feedback (by copying the initial weights) of the diffusion network it aims to control, whereas direct

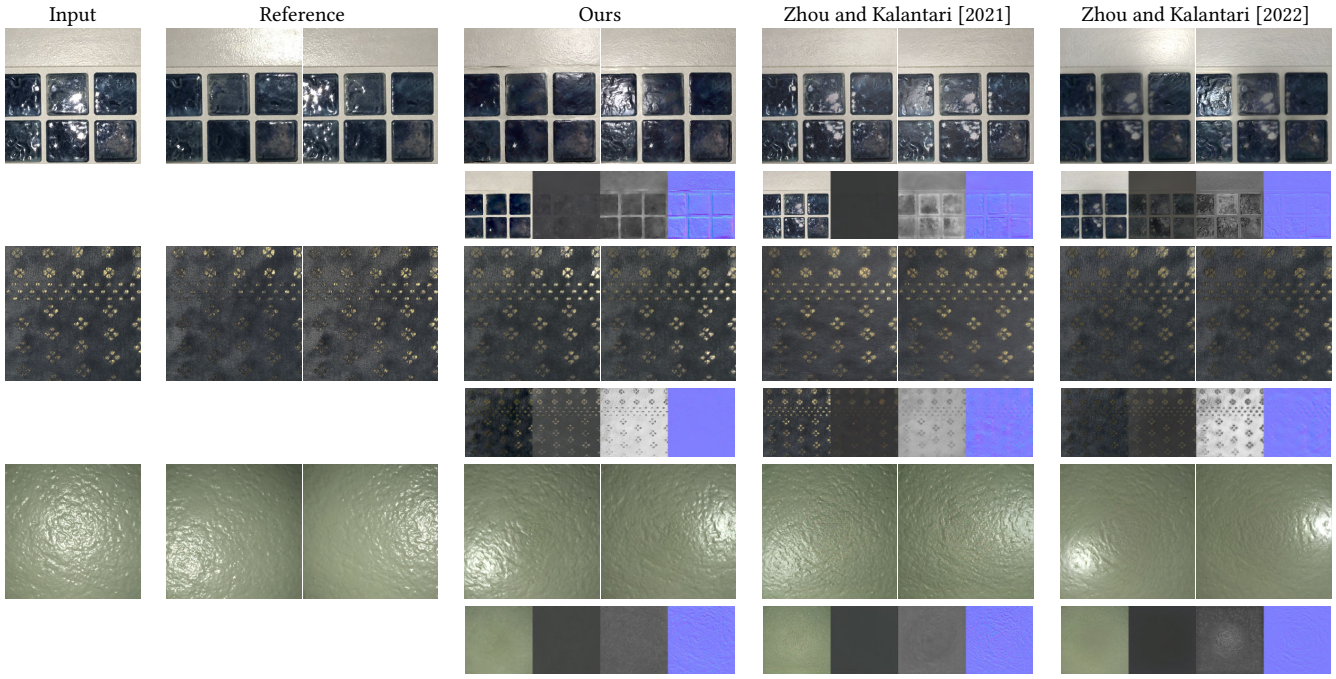

Figure 6: Qualitative comparison on real-world materials captured with a colocated light source, and relit from two different point light positions.

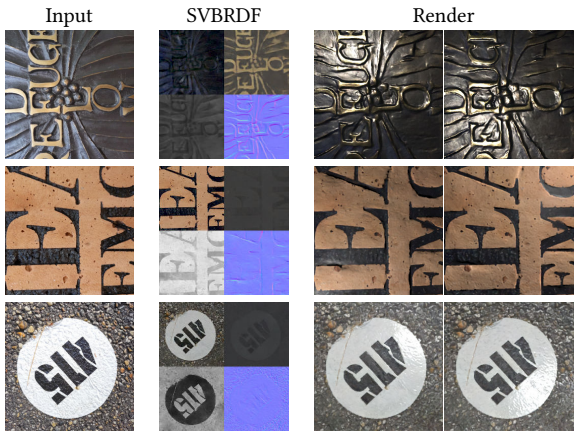

Figure 7: Demonstration of in-the-wild SVBRDF capture under uncontrolled unknown natural lighting and revisualized under novel lighting.

conditioning closely intertwines both control and synthesis. Furthermore, our input conditions are more strict, leaving less room for synthesis than typical ControlNet conditions (e.g., sketches). However, our conclusions with respect to ControlNet are only validated for MatFusion using photographs as conditions, and further investigations are needed to ascertain whether these conclusions extend to other diffusion networks and/or condition types.

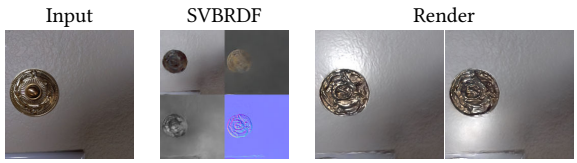

Figure 8: Failure case: artificial "blob-like" normal maps.

Limitations. MatFusion is a generative SVBRDF model, and it has trouble generating pixel-perfect reproductions. Hence, MatFusion does not necessary produce the lowest errors on pixel-based metrics. Furthermore, as a generative model, MatFusion is better suited for capturing materials with organic structures than those with regular straight lines. We posit that this is the reason why MatFusion tends to produce higher quality results on real-world captures than on artist-generated materials which are more regular. This causes MatFusion to sometimes generate properties maps that look too artificial (Figure 8). Furthermore, MatFusion is currently limited to 256×256 resolution SVBRDFs. Finally, the render error selection requires prior knowledge of the lighting condition, hampering automatic selection from photographs under unknown lighting (e.g., natural lighting). Furthermore, it does not always yield a good selection because oversaturation can make it difficult to differentiate between two SVBRDFs that produce a similar rendered replica but that substantially differ in quality. Ideally, we would like to employ a selection criterion that judges plausibility of the SVBRDFs.

6 CONCLUSION

We presented MatFusion, a generative SVBRDF diffusion model trained on a new large and diverse training set of synthetic SVBRDFs. MatFusion can subsequently serve as a starting point for refining an SVBRDF diffusion model conditioned on captured images under some target lighting condition. We demonstrated the flexibility and efficacy of MatFusion by training three conditional variants: one for photographs captured with a colocated flash light, one under unknown and uncontrolled natural lighting, and one for flash/no-flash image pairs. An advantage of using a generative SVBRDF model is that different replicates can be synthesized by changing the seed, allowing user to select the most plausible replicate. For future work we would like to investigate more comprehensive metrics for automatic selection, and better regularization during training and/or inference for modeling regular features. Based on the recent successes in coupling large language models with diffusion models, another interesting avenue would be to explore better authoring tools for SVBRDF creation.

ACKNOWLEDGMENTS

This research was supported in part by NSF grant IIS-1909028.

REFERENCES

Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2016. Reflectance modeling by neural texture synthesis. *ACM Trans. Graph.* 35, 4 (2016).

Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2018. Single-image SVBRDF capture with a rendering-aware deep network. *ACM Trans. Graph.* 37, 4 (2018).

Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2019. Flexible SVBRDF Capture with a Multi-Image Deep Network. *Comp. Graph. Forum* 38, 4 (2019).

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, Vol. 34. 8780–8794.

Michael Fischer and Tobias Ritschel. 2022. Metappearance: Meta-Learning for Visual Appearance Reproduction. *ACM Trans. Graph.* 41, 6, Article 245 (nov 2022).

Duan Gao, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2019. Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. *ACM Trans. Graph.* 38, 4 (2019).

Paul Guerrero, Miloš Hašan, Kalyan Sunkavalli, Radomír Měch, Tamy Boubekeur, and Niloy J. Mitra. 2022. MatFormer: A Generative Model for Procedural Materials. *ACM Trans. Graph.* 41, 4, Article 46 (jul 2022).

Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan. 2021. Highlight-Aware Two-Stream Network for Single-Image SVBRDF Acquisition. *ACM Trans. Graph.* 40, 4, Article 123 (2021).

Y. Guo, M. Hašan, L. Yan, and S. Zhao. 2020a. A Bayesian Inference Framework for Procedural Material Parameter Estimation. *Comp. Graph. Forum* 39, 7 (2020), 255–266.

Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. 2020b. MaterialGAN: Reflectance Capture Using a Generative SVBRDF Model. *ACM Trans. Graph.* 39, 6, Article 254 (2020).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. 2023. *Flax: A neural network library and ecosystem for JAX.* http://github.com/google/flax

Philipp Henzler, Valentin Deschaintre, Niloy J. Mitra, and Tobias Ritschel. 2021. Generative Modelling of BRDF Textures from Flash Images. *ACM Trans. Graph.* 40, 6, Article 284 (2021).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *J. Mach. Learn. Res.* 23 (2022), 47:1–47:33.

Yiwei Hu, Miloš Hašan, Paul Guerrero, Holly Rushmeier, and Valentin Deschaintre. 2022a. Controlling Material Appearance by Examples. *Comp. Graph. Forum* 41, 4 (2022), 117–128.

Yiwei Hu, Chengan He, Valentin Deschaintre, Julie Dorsey, and Holly Rushmeier. 2022b. An Inverse Procedural Modeling Pipeline for SVBRDF Maps. *ACM Trans. Graph.* 41, 2, Article 18 (jan 2022).

Zahra Kadkhodaie and Eero Simoncelli. 2021. Stochastic Solutions for Linear Inverse Problems using the Prior Implicit in a Denoiser. In *NeurIPS*, Vol. 34. 13242–13254.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *NeurIPS*.

Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Trans. Graph.* 36, 4 (2017).

Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. 2018. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. In *ECCV*. 74–90.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. *CVPR* (2022).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Rosalie Martin, Arthur Roullier, Romain Rouffet, Adrien Kaiser, and Tamy Boubekeur. 2022. MaterIA: Single Image High-Resolution Material Capture in the Wild. *Comp. Graph. Forum* 41, 2 (2022), 163–177.

Yvain Quéau, Jean-Denis Durou, and Jean-Franccois Aujol. 2018. Normal Integration: A Survey. *Journal of Mathematical Imaging and Vision* 60, 4 (May 2018), 576–593.

Markus N. Rabe and Charles Staats. 2021. Self-attention Does Not Need $O(n^2)$ Memory. arXiv:2112.05682 [cs.LG]

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*. 10684–10695.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-Image Diffusion Models. In *ACM SIGGRAPH 2022 Conference Proceedings (SIGGRAPH '22)*. Article 15, 10 pages.

C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. 2023. Image Super-Resolution via Iterative Refinement. *IEEE TPAMI* 45, 04 (apr 2023), 4713–4726.

Shen Sang and M. Chandraker. 2020. Single-Shot Neural Relighting and SVBRDF Estimation. In *ECCV*.

Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. 2021. UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models. arXiv:2104.05358 [cs.CV]

Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik. 2020. Match: Differentiable Material Graphs for Procedural Material Capture. *ACM Trans. Graph.* 39, 6, Article 196 (nov 2020).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021a. Denoising Diffusion Implicit Models. In *ICLR*.

Yang Song and Stefano Ermon. 2020. Improved techniques for training score-based generative models. *NeurIPS* 33 (2020), 12438–12448.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021b. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

Giuseppe Vecchio, Simone Palazzo, and Concetto Spampinato. 2021. SurfaceNet: Adversarial SVBRDF Estimation From a Single Image. In *ICCV*.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Andrey Voynov, Kfir Abernan, and Daniel Cohen-Or. 2022. Sketch-Guided Text-to-Image Diffusion Models. (2022).

Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. 2007. Microfacet Models for Refraction through Rough Surfaces. In *EGSR*. 195–206.

Tao Wen, Beibei Wang, Lei Zhang, Jie Guo, and Nicolas Holzschuch. 2022. SVBRDF Recovery from a Single Image with Highlights Using a Pre-trained Generative Adversarial Network. *Comp. Graph. Forum* 41, 6 (2022).

Wenjie Ye, Yue Dong, Pieter Peers, and Baining Guo. 2021. Deep Reflectance Scanning: Recovering Spatially-varying Material Appearance from a Flash-lit Video Sequence. *Comp. Graph. Forum* 40, 6 (2021), 409–427.

Wenjie Ye, Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2018. Single Image Surface Appearance Modeling with Self-augmented CNNs and Inexact Supervision. *Comp. Graph. Forum* 37, 7 (2018), 201–211.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs].

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. 2022. TileGen: Tileable, Controllable Material Generation and Capture. In *SIGGRAPH Asia 2022 Conference Papers*. Article 34.

Xilong Zhou and Nima Khademi Kalantari. 2021. Adversarial Single-Image SVBRDF Estimation with Hybrid Training. *Comp. Graph. Forum* (2021).

Xilong Zhou and Nima Khademi Kalantari. 2022. Look-Ahead Training with Learned Reflectance Loss for Single-Image SVBRDF Estimation. *ACM Trans. Graph.* 41, 6, Article 266 (nov 2022).

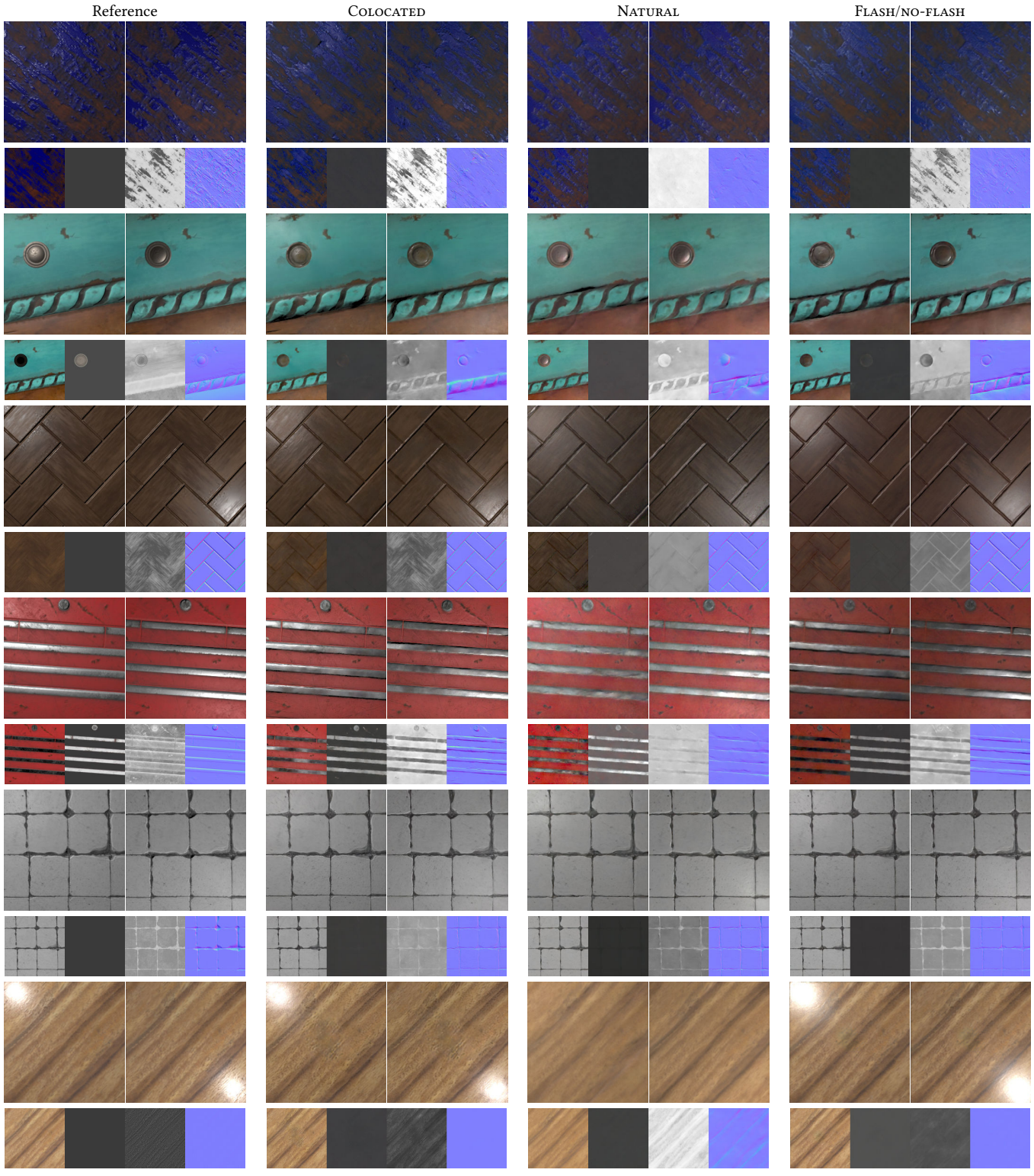

Figure 9: Comparison of the Colocated, Natural, and Flash/no-flash conditional diffusion models on a variety of synthetic SVBRDFs.

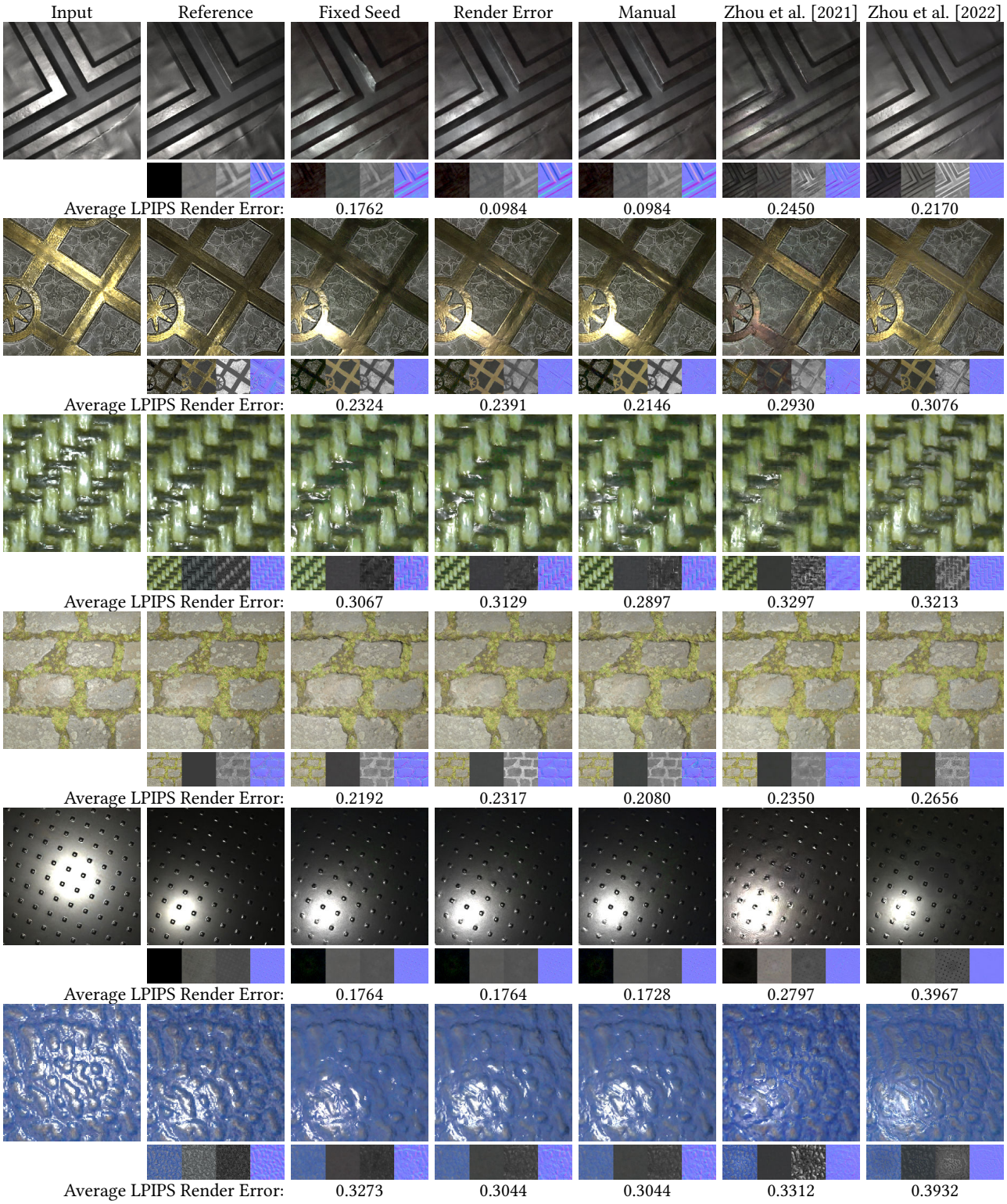

Figure 10: Qualitative comparison of MatFusion conditioned on colocated lighting (*fixed seed*, *render error*, and *manual* selection) against the adversarial direct inference of Zhou and Kalantari [2021] and the meta-leanring look-ahead method of Zhou and Kalantari [2022]. The LPIPS errors are averaged over visualizations under 128 different point lights sampled on the hemisphere surrounding the sample.